O.S.P.

L'orientation scolaire et professionnelle

43/1 | 2014
Varia

L'analyse des erreurs dans les tests de raisonnement logique : principes et illustrations

The error analysis in the tests of logical reasoning: principles and illustrations

Philippe Chartier, Hana Barbot et Rodrigue Ozenne

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/osp/4296
DOI : 10.4000/osp.4296
ISSN : 2104-3795

Éditeur
Institut national d'étude du travail et d'orientation professionnelle (INETOP)

Édition imprimée
Date de publication : 7 mars 2014
ISSN : 0249-6739

Référence électronique
Philippe Chartier, Hana Barbot et Rodrigue Ozenne, « L'analyse des erreurs dans les tests de raisonnement logique : principes et illustrations », *L'orientation scolaire et professionnelle* [En ligne], 43/1 | 2014, mis en ligne le 07 mars 2017, consulté le 10 décembre 2020. URL : http://journals.openedition.org/osp/4296 ; DOI : https://doi.org/10.4000/osp.4296

L'analyse des erreurs dans les tests de raisonnement logique : principes et illustrations

The error analysis in the tests of logical reasoning: principles and illustrations

Philippe Chartier, Hana Barbot et Rodrigue Ozenne

1 Dans les tests de raisonnement logique (épreuves d'intelligence générale de type facteur g), le psychologue ne prend généralement en compte qu'un seul indicateur de la performance du sujet : son score total. Cette centration sur le niveau de performance à l'épreuve apporte peu d'informations sur les conditions de réalisation de cette performance, limite souvent soulignée de ce type de test (Chartier & Loarer, 2008 ; Grégoire, 2004 ; Hessels & Hessels-Schlatter, 2010 ; Huteau & Lautrey, 1999). D'autres variables mériteraient pourtant une attention particulière, car elles peuvent apporter des pistes explicatives à ces performances. C'est par exemple le cas des stratégies de résolution, aspect régulièrement abordé dans les publications relevant de la psychologie différentielle (e.g. de Ribaupierre, Ghisletta, Lecerf, & Roulin, 2010) ou encore les erreurs, aspect beaucoup moins fréquent dans les analyses de protocoles. L'objectif de cet article, à caractère exploratoire, est de proposer un cadre général d'analyse des erreurs commises par les sujets dans des tests de raisonnement et de l'illustrer à travers deux épreuves : les matrices de Raven et le test RCC (Chartier, 2012).

L'analyse de l'erreur

2 Cette approche a été surtout développée dans les sciences de l'éducation et la didactique des disciplines qui ont donné un nouveau statut à l'erreur, dans le cadre d'une évaluation formative, ou diagnostique, des apprentissages : « on est globalement passé d'une conception négative donnant lieu à une sanction à une autre, où les erreurs se présentent plutôt comme des indices pour comprendre le processus d'apprentissage et comme

témoins pour repérer les difficultés des élèves » (Astolfi, 1997). À partir de cette rupture épistémologique (Favre, 1995), l'erreur n'est plus une faute (Favre, 2010), mais devient un indicateur, une information pertinente, un objet possible d'analyse. Il s'agit, par exemple, de rendre compte des procédures de résolution utilisées par un élève qui les a amenées à fournir une réponse fausse, de retrouver les opérations intellectuelles dont elles sont la trace ou encore de renseigner sur le type de représentation mentale de la situation élaboré par une personne (Astolfi, 1997). On retrouve un intérêt pour l'analyse des erreurs dans le cadre plus général de la psychologie cognitive (Fayol, 1995) ou encore en ergonomie (Leplat, 1999) mais, à notre connaissance, cet aspect semble assez négligé en psychométrie, et tout parti-culièrement dans les tests de raisonnement.

L'analyse des erreurs dans les tests de raisonnement

3 Dans ce type d'épreuve, la pratique de cotation la plus fréquente consiste à identifier une (et une seule) bonne réponse pour chaque item, à accorder un point, puis à additionner les points obtenus pour calculer le score total de l'épreuve. L'évaluation ne repose donc que sur le nombre d'items réussis sans tenir compte des autres informations présentes dans le protocole. On considère alors, au moins de manière implicite, qu'il n'existe pas de différences interindividuelles, ou du moins de différences interindividuelles méritant d'être prises en compte, autres que le score total. Pourtant, on peut envisager que des sujets obtenant un même score à une épreuve puissent se différencier sur les items qu'ils ont réussis (leur profil de bonnes réponses)[1] et que des sujets échouant à un même item ne fournissent pas obligatoirement la même réponse erronée (le type d'erreur).

4 Nous n'aborderons ici que la seconde forme de différence interindividuelle, liée aux erreurs commises (pour une illustration de la première forme, voir Chartier, 2010). Comme nous l'avons signalé, les publications disponibles sur cet aspect sont assez rares. L'un des seuls exemples disponibles concerne les matrices de Raven[2]. Dans l'ouvrage introductif de ces épreuves, les auteurs précisent que, même si l'épreuve n'a pas été conçue dans cet objectif, l'analyse des erreurs peut être source d'information : « Dans les Matrices, les dessins constituant une réponse fausse n'ont pas été choisis pour prêter délibérément à confusion, ni pour révéler quoi que ce soit sur les types d'erreurs faites par le sujet. Néanmoins, le type de mauvaises réponses indiquera parfois à quoi une personne échoue et, dans une certaine mesure, pour quelles raisons. » (Raven, Raven & Court, 1998, section 1, p. 53). Pour compléter leur propos, ces mêmes auteurs fournissent, dans la version APM[3] de leur épreuve, une typologie des erreurs qui distingue quatre catégories (Raven, Raven, & Court, 1998, pp. 15-19) :

5 - solution incomplète (type A) : identification incomplète des règles à prendre en compte ;

6 - mode de raisonnement arbitraire (type B) : principe de raisonnement qualitativement différent de celui que requiert l'item ;

7 - choix surdéterminé par des éléments intrus (type C) : solutions qui combinent le maximum d'éléments disponibles ;

8 - répétitions (type D) : sélection d'une figure identique à l'une des figures du problème.

9 Ces types d'erreurs se différencient au niveau de leur qualité : certaines étant proches de la bonne réponse (type A), d'autres beaucoup plus éloignées (type B ou D). On observe que la proportion de ces différentes erreurs varie selon le niveau de compétence des sujets : les sujets de faible niveau commettent plus d'erreurs de type B alors que ce sont les

erreurs de type A qui sont les plus fréquentes chez les sujets de niveau de compétence plus élevé. De même, pour Vodegel-Matzen, Van der Molen et Dudink (1994), le type d'erreurs commises varie en fonction du score total au test, résultats similaires observés également par Babcock (2002). Si ces erreurs sont de qualité différente, pourquoi alors ne pas en tenir compte dans la cotation de l'épreuve ? Une possibilité de cotation plus fine, qui sortirait de la conception binaire classique juste/faux, est évoquée d'ailleurs dans le manuel de la version standard du Raven (SPM[4]) qui envisage une pondération de la cotation avec la possibilité d'une « approche plus fine, avec un éventail de notes plus étendu que la simple notation juste/faux traditionnelle » (Raven, Court & Raven, 1998, p. 36). Mais les auteurs ne poursuivent pas plus loin cette possibilité et, à notre connaissance, cette pondération de la cotation ne semble pas avoir été développée.

10 On peut retrouver une finesse d'analyse des réponses erronées dans certains subtests des échelles de Wechsler pour lesquels trois niveaux de cotation sont envisagés en fonction de la qualité de la réponse : 0, 1 ou 2 points (Wechsler, 2005). Dans ces tests, on peut observer qu'il existe un échelon entre la réponse fausse (notée 0) et la bonne réponse (notée 2) : une réponse à 1 point n'est pas une réponse complètement fausse, mais elle n'est pas non plus entièrement juste.

11 Plus récemment, Loye (2010), dans sa présentation des modèles de classification diagnostique, aborde la possibilité, dans ces modèles, de prévoir des variables observées polychotomiques qui distinguent plusieurs niveaux de réussite. On retrouve également ce principe dans la cotation de certains items de l'enquête PISA pour lesquels il est prévu une cotation selon une modélisation à crédit partiel à trois niveaux : crédit complet, crédit partiel et absence de crédit (Lafontaine & Raîche, 2011). Ici encore est envisagé un échelon intermédiaire entre la mauvaise réponse (absence de crédit) et la bonne réponse (crédit complet).

12 Enfin, signalons qu'une analyse des erreurs a également été menée pour le test BLS4 dans l'objectif d'en identifier les régularités (Thiébaut, 2000). Le manuel de ce test répertorie les erreurs les plus fréquentes pour les 16 premiers items de l'épreuve, mais l'analyse reste assez descriptive et ne permet pas d'identifier des différences de qualité dans les réponses fausses observées.

13 On peut également citer ici la notion de test différentiel, développée par Bonnardel dans les années cinquante, qui se caractérise par différentes possibilités de réponse à un item, chacune reflétant un niveau de compétence identifié au préalable (Bonnardel, 1957). Par exemple, sur le test BV16, quatre catégories de réponses sont distinguées : « bonnes réponses » (cotées + 2), « réponses médiocres » (0), « mauvaises réponses » (– 1) et « très mauvaises réponses » (– 2). Sans remettre en cause l'intérêt de cette approche, Thiébaut (2010) ne vérifie pas les conditions de validité de ce type de test sur des données récentes du BV16.

14 De toutes ces recherches, nous pouvons retenir qu'il semble possible d'envi-sager une cotation plus fine des protocoles des sujets, après identification de différentes qualités de réponse, pour dépasser la cotation binaire classique juste/faux, 0/1, et envisager alors une cotation plus détaillée, apportant davantage d'informations sur la performance observée. Nous allons maintenant présenter comment nous avons développé ces principes à partir de deux épreuves.

Nos propositions sur l'analyse des erreurs

15 Nous présentons ici, pour chacune des épreuves utilisées, dans un premier temps, l'approche descriptive qui consiste à identifier différents types d'erreurs, et dans un second temps une approche plus quantitative qui propose le calcul d'indicateurs.

L'analyse des erreurs dans les matrices de Raven (SPM)

16 Le SPM permet, selon son auteur, de mesurer l'aptitude actuelle d'un sujet à percevoir et à penser clairement. Ce test de raisonnement est reconnu comme étant l'une des meilleures estimations du facteur g (Huteau & Lautrey, 1999). Le SPM est composé de cinq séries (A, B, C, D, E) de douze items numérotés de 1 à 12, où le sujet est invité à trouver, parmi six ou huit possibilités, la pièce manquante pour compléter la suite logique. Chaque série commence par un problème dont la résolution est facile, introduisant ainsi une manière de raisonner qui va se complexifier progressivement. De plus, les séries sont elles aussi de plus en plus difficiles, ainsi la série A mesure l'aptitude des sujets à compléter des patterns continus, la série B contient des items dont la solution est obtenue par raisonnement analogique, les derniers items de cette série étant du même ordre de difficulté que les premiers des trois séries suivantes.

17 Comme nous venons de le voir, des tentatives d'analyse et de classification des erreurs sont mentionnées dans les manuels des APM, mais rien n'est précisé dans le manuel des SPM. Après avoir procédé à une analyse des cinq séries d'items de cette version SPM (voir annexe 1), nous avons cherché à catégoriser les distracteurs proposés dans cette épreuve. À partir des propositions de ces manuels et d'un travail qui avait été mené par l'équipe du CIO spécialisé de Lille dans les années quatre-vingt-dix (Bellina, non daté), nous proposons la typologie suivante des distracteurs de la version SPM des Raven (voir en annexe 2 la catégorisation de toutes les réponses) :

18 - erreur de type 1 : choix de voisinage, réponse identique à la forme qui se trouve à proximité ;

19 - erreur de type 2 : un élément sur deux est juste ; le sujet fait preuve d'un début de raisonnement, mais ne peut pas prendre en considération la totalité des éléments à analyser ;

20 - erreur de type 3 : erreur d'orientation ; la réponse est juste mais inversée dans l'espace gauche/droite ou haut/bas ;

21 - erreur de type 4 : tous les éléments de la figure sont rassemblés dans la réponse proposée ;

22 - erreur de type 5 : réponse aberrante, aucun repère de raisonnement n'est identifiable.

23 Comme pour les autres formes des matrices, les erreurs ne sont pas de même niveau et diffèrent quant à leur proximité avec la bonne réponse. On considère que :

24 - les erreurs de type 1 et 5 sont de « vraies erreurs » : rien dans la réponse donnée n'indique un raisonnement même partiel de la part du sujet ;

25 - les erreurs de type 4 indiquent que le sujet n'arrive pas à sélectionner les indices pertinents pour donner sa réponse : il choisit la solution qui présente le maximum d'éléments ;

26 - les erreurs de type 2 et 3 sont des réponses partiellement justes : le raisonnement est correct, mais ne prend pas en compte toutes les données du problème.

27 Au niveau descriptif, il est alors possible d'observer quelles sont les types d'erreurs les plus fréquentes chez les personnes afin de mieux comprendre les difficultés rencontrées dans l'épreuve.

28 Dans une approche plus quantitative, nous avons élaboré deux indicateurs à partir de cette typologie. Nous proposons de considérer les réponses de type 2 et 3 comme des bonnes réponses partielles (BRP), car elles sont plus proches de la bonne réponse que les autres types d'erreurs. Mais comment tenir compte de ces différences dans un score ? Nous avons décidé d'accorder ½ point[5] à chaque réponse de ce type (rappelons qu'une bonne réponse est cotée 1 point). En reprenant les réponses fausses et en accordant donc ½ point à chaque réponse de type BRP on obtient le score de Réussite Partielle (RP). En ajoutant ce score RP au score Raven initial, on obtient le score Raven potentiel. Au final, deux indicateurs de la performance de chaque sujet sont alors disponibles : son score Raven classique, son score Raven potentiel.

L'analyse des erreurs dans le RCC

29 Ce test vise à évaluer les capacités de raisonnement logique à partir d'un support original : des cartes à jouer (Chartier, 2008, 2010). La tâche est inspirée du test MGM de Pire (1957) dans lequel le sujet doit découvrir les caractéristiques (famille et valeur) d'une carte retournée qui complète une série proposée. Pour les familles, il existe quatre possibilités de réponse (carreau, cœur, trèfle et pique), pour les valeurs dix possibilités (seules les cartes de 1 à 10 sont utilisées). Le principe de cotation est le suivant : la réponse est considérée comme juste (1 point) si la personne fournit à la fois la bonne famille et la bonne valeur. En additionnant les items réussis, on obtient le score RCC.

30 Ce principe général de cotation peut être discuté, car il revient à attribuer la note 0 à toute réponse différente de la bonne réponse attendue, alors même qu'il peut exister des différences de qualité entre les réponses fausses. Ainsi, il nous semble pertinent de repérer ici au moins deux types d'erreurs :

31 - type 1 : des réponses entièrement fausses qui ne comportent aucun élément de la bonne réponse (valeur et famille fausses) ;

32 - type 2 : des réponses erronées qui comportent une partie de la bonne réponse, c'est-à-dire un élément correct (valeur ou famille) parmi les deux attendus.

33 En adoptant la même logique que celle appliquée aux matrices de Raven, nous proposons de nommer ces réponses de type 2 bonnes réponses partielles (BRP), car elles sont plus proches des bonnes réponses attendues que les réponses de type 1 et comportent de fait une facette du raisonnement que l'on cherche à estimer par le test RCC. Deux types de BRP peuvent être distingués ici :

34 - des réponses qui comportent la bonne valeur, mais avec erreur pour la famille, nous nommerons ces réponses des bonnes réponses partielles valeur (BRP valeur) ;

35 - le cas inverse : des réponses qui comportent la bonne famille, mais avec erreur sur la valeur, nous nommerons ces réponses des bonnes réponses partielles famille (BRP famille).

36 Dans une approche descriptive, nous pouvons ainsi repérer, par exemple, quelles sont les types d'erreurs les plus fréquentes pour un individu donné. Mais l'analyse peut se poursuivre par une approche quantitative et le calcul de deux indicateurs, en suivant toujours la même logique que celle proposée pour les matrices de Raven : en reprenant les réponses fausses et en accordant ½ point[6] aux réponses BRP, on obtient le score de réussite partielle (RP) ; en ajoutant ce score RP au score RCC initial, on obtient le score RCC potentiel. Comme pour les matrices, au final le sujet peut être caractérisé par deux indicateurs : son score RCC et son score RCC potentiel.

Problématique et méthodologie

37 La problématique générale concerne la vérification de l'existence de différences entre des sujets sur la qualité des erreurs commises dans les épreuves proposées. Deux analyses peuvent être menées, l'une reposant sur des données descriptives, la seconde sur des données quantitatives et le calcul d'indicateurs.

Participants

38 Il s'agit d'élèves scolarisés en classe de 6e (âge moyen : 11 ans 9 mois) dans deux collèges de la région parisienne. Cent quatre-vingt-trois ont passé l'épreuve matrices SPM (101 filles et 82 garçons) et 123 l'épreuve RCC (71 filles et 52 garçons).

Procédure

39 L'épreuve matrices de Raven version SPM, composée de 60 items, a été passée selon les conditions de standardisation, en temps limité de 20 minutes. L'épreuve RCC, qui comporte 40 items, a été passée également en temps limité de 20 minutes.

Résultats

Résultats sur l'épreuve matrices SPM de Raven

40 Les données descriptives du score aux matrices SPM figurent dans le tableau 1. Les valeurs nous indiquent que le test présente des qualités métriques satisfaisantes tant au niveau de la sensibilité que de son homogénéité. Avec un score moyen de 37 points, les élèves fournissent en moyenne 23 réponses fausses ou absence de réponse.

Table 1

Statistiques descriptives du score brut total aux matrices de Raven

	Nombre d'items	M	ET	Min	Max	alpha de Cronbach
Score Raven SPM	60	37.01	8.7	9	54	.90

Table 1
Descriptive statistics of the total raw score on Raven's Progressive Matrices

41 Si nous appliquons nos propositions de classification des erreurs en cinq catégories, nous obtenons les résultats suivants (tableau 2).

42 On peut observer que les réponses les plus fréquentes sont de type 1 (choix de voisinage) et de type 5 (réponse aberrante), avec en moyenne plus de cinq réponses par élève de chacun de ces deux types d'erreurs, les plus rares étant les réponses de type 3 (bonne réponse, mais mal orientée) avec en moyenne un peu plus d'une réponse de ce type. Si on analyse maintenant les deux réponses considérées comme étant de qualité supérieure, réponses de type BRP (T2 et T3), on peut noter qu'il y a trois fois plus de T2 (un élément correct dans la réponse) que de T3. Mais attention ici, car la répartition de chaque catégorie de distracteurs n'étant pas identique pour tous les items (voir en annexe 2 le détail), la comparaison des moyennes n'a guère de sens. Par contre, ce que nous indique ce tableau, c'est l'existence d'une grande variabilité interélèves sur la qualité des réponses fausses : certains ne fournissent jamais de réponse de type BRP, tandis que pour d'autres ces réponses sont très fréquentes (voir les valeurs minimales et maximales de T2 et T3 dans le tableau 2).

43 En accordant ½ point pour chaque réponse BRP, nous pouvons calculer nos indicateurs RP et Raven potentiel (voir tableau 3).

Table 2

Statistiques descriptives des réponses fausses aux matrices de Raven par types d'erreurs

Type d'erreur	M	ET	Min	Max
T1	5.70	5.15	0	30
T2	4.04	2.59	0	13
T3	1.34	.94	0	5
T4	2.81	2.06	0	10
T5	5.29	3.07	0	14

Table 2
Descriptive statistics of wrong answers on Raven's Progressive Matrices by error type

Table 3

Statistiques descriptives pour le nombre de RP et le score potentiel aux matrices de Raven

	M	ET	Min	Max
Réussites partielles (RP)	2.69	1.54	0	8
Score Raven potentiel	39.70	7.72	16	55

Table 3
Descriptive statistics for RP number and potential score on Raven Matrices

44 En moyenne, les élèves obtiennent 2,69 points pour le score RP, soit un peu plus de cinq réponses de type BRP. Même si la corrélation entre les scores Raven et Raven potentiel est très élevée[7] (r = .99), on observe des écarts individuels non négligeables, car le score RP varie de 0 à 8 points. On retrouve ici les constats du tableau précédent : les élèves peuvent se différencier au niveau de la qualité des erreurs commises. Il est difficile d'interpréter ces résultats, car le niveau de réussite global a un effet sur la possibilité concrète d'obtenir des points supplémentaires (pour un score Raven élevé, le nombre d'erreurs est faible, donc la marge possible de progression du score RP est beaucoup plus réduite que dans le cas d'un score Raven plus faible et donc un nombre d'erreurs plus élevé). Il semble plus intéressant alors de comparer des élèves ayant le même score Raven. Pour effectuer cette comparaison, nous avons sélectionné les scores Raven les plus fréquents et avons relevé les scores Raven potentiel des élèves concernés (voir tableau 4).

Table 4

Variabilité des scores RP et Raven potentiel pour les scores Raven les plus fréquents

	Scores Raven				
	37	39	40	41	43
Effectif	12.0	13.0	12.0	12.0	12.0
Score RP min	1.5	1.0	1.0	1.5	0.5
Score RP max	4.0	4.5	33.0	4.0	3.0
Variation	2.5	3.5	2.0	3.5	2.5
Score Raven potentiel	38.5 à 41	40 à 43.5	41 à 43	42.5 à 45	43.5 à 46

Table 4
Variability in RP and potential Raven scores for the most frequent Raven scores

45 Pour tous les scores Raven sélectionnés, on observe des variations sur les scores Raven potentiel comprises entre 2 et 3,5 points. Par exemple, pour un score Raven de 39 points, le score Raven potentiel varie de 40 à 43,5 points. Des élèves considérés comme

comparables au niveau de leur capacité générale de raisonnement, si l'on tient compte de leur score Raven, peuvent ainsi être différenciés si l'on prend en compte la qualité de leurs réponses fausses (leur score Raven potentiel).

Résultats sur l'épreuve RCC

46 Les données descriptives du score RCC figurent dans le tableau 5. Les valeurs nous indiquent que le test présente des qualités métriques satisfaisantes tant au niveau de la sensibilité que de son homogénéité (coefficient alpha de Cronbach de 0,87). Comme attendu, les élèves se différencient sur le score RCC qui témoigne de leur compétence générale de raisonnement logique. Rappelons que ce score RCC ne prend en compte que les réponses complètement correctes (famille et valeur justes).

Tableau 5

Statistiques descriptives au RCC

	Nombre d'items	M	ET	Min	Max	alpha de Cronbach
Score CC	40	18.6	6.6	1	33	.87

Table 5
Descriptive statistics for RCC

47 En reprenant notre cadre d'analyse des erreurs, qui distingue les réponses entièrement fausses (type 1) des réponses de type BRP, nous pouvons analyser la répartition de ces réponses dans le tableau suivant :

Table 6

Taux de réponses fausses au RCC (en %)

	Proportion	Min	Max
T1	82	50	100
T2 : BRP valeur	6	0	27
T2 : BRP famille	12	0	40

Table 6
Rate of false responses in RCC (in %)

48 On peut observer que les réponses de type 1 sont les plus fréquentes, les réponses de type 2 ne représentant que 18 % des réponses fausses. Parmi celles-ci, les BRP famille sont deux fois plus nombreuses que les BRP valeur, ce qui n'est pas étonnant, car il est statistiquement plus probable de trouver la bonne famille que de trouver la bonne valeur (il n'existe seulement que quatre possibilités de familles contre dix possibilités de valeurs). Mais ce que ce tableau nous apporte, c'est l'illustration d'une variabilité

importante entre les élèves : certains ne fournissant aucune réponse de type 2, alors que pour d'autres élèves ce type de réponse est fréquent et peut atteindre 27 % pour les BRP valeur et 40 % pour les BRP familles. On retrouve ici des différences interindividuelles concernant la qualité des réponses erronées.

49 L'étape suivante consiste à dépasser ce constat descriptif pour aborder une analyse quantitative à partir de nos deux indicateurs (voir tableau 7).

Tableau 7

Statistiques descriptives pour le nombre de RP et le score potentiel au RCC

	M	ET	Min	Max
RP	3.6	2.5	1.0	14.5
RCC potentiel	22.2	6.3	6.5	36.5

Table 7
Descriptive statistics for RP number and potential RCC score

50 Comme pour le Raven, même si la corrélation entre RCC et RCC potentiel est très élevée (r = .92), des différences interindividuelles non négligeables apparaissent ici sur le score RP : les élèves obtiennent en moyenne 3,6 points, mais on retrouve les écarts individuels constatés précédemment, car ce score RP varie de 1 à 14,5 points. On retrouve logiquement ces résultats sur le score RCC potentiel.

51 En comparant les scores RCC et RCC potentiel (tableaux 5 et 7), on observe que l'augmentation est plus importante pour les scores RCC faibles (passage du score minimum de 1 à 6,5) que pour les scores élevés (passage de 33 à 36,5). Effectivement, comme nous l'avons déjà indiqué, il est plus facile d'obtenir des demi-points supplémentaires lorsque le protocole comporte beaucoup d'items échoués que lorsque le score est proche du maximum. Pour annuler les effets possibles du niveau de performance sur les différences individuelles en matière de qualité d'erreur, il faut donc effectuer les comparaisons à score RCC constant. Pour cette raison, nous avons sélectionné les scores RCC les plus fréquents et avons relevé les scores RCC potentiel des élèves concernés (voir tableau 8). La question principale est la suivante : à niveau équivalent, les élèves peuvent-ils être différenciés sur le score potentiel ?

Tableau 8

Variabilité des scores RP et RCC potentiel pour les scores RCC les plus fréquents

		Scores RCC		
	14	16	22	23
Effectif	11.0	9.0	13.0	8.0
Moyenne RP	3.1	4.6	3.3	4.4

Score RP min	0.5	1.0	0.5	1.0
Score RP max	8.0	10.0	7.5	6.5
Variation RP	7.5	9.0	7.0	5.5
Score RCC potentiel	14.5 à 22	17 à 26	22.5 à 29.5	24 à 29.5

Table 8
Variability in RP and potential RCC scores for the most frequent RCC scores

52 Quel que soit le score RCC d'origine (ici 14, 16, 22 ou 23), on observe des différences interélèves importantes sur les scores RCC potentiel correspondants, avec des variations comprises entre 5,5 points et 9 points. Par exemple, pour un score RCC de 16 points, le score RCC potentiel varie de 17 à 26 points. Ainsi, certains élèves, lorsqu'ils échouent, produisent souvent des erreurs de type 1 (ce qui aboutit alors à un score RP faible et aux scores RCC potentiel les plus faibles du sous-groupe), alors que d'autres élèves, de même niveau global de raisonnement[8], fournissent plus fréquemment des réponses fausses de type 2, plus proches des bonnes réponses attendues (d'où un score RP élevé, qui aboutit aux scores RCC potentiel les plus élevés du sous-groupe). Et ces différences se retrouvent à tous les niveaux de réussite. Autrement dit, des élèves considérés comme comparables vis-à-vis de leur compétence générale de raisonnement (scores RCC identiques) peuvent présenter des différences parfois importantes lorsque l'on prend en compte la qualité de leurs réponses fausses.

53 Une autre manière d'analyser ces différences est d'observer, pour un même score RCC potentiel, les scores RCC initiaux. Pour illustrer cet aspect, nous avons sélectionné les deux scores RCC potentiel les plus fréquents.

54 Ici encore, nous observons des différences individuelles non négligeables : pour le même score RCC potentiel, les élèves peuvent avoir un score initial RCC très différent. Par exemple pour le score RCC potentiel de 26, le score RCC initial varie entre 16 et 24, soit un niveau de performance très différent si l'on se réfère au seul score RCC.

Tableau 9

Variabilité des scores RCC pour les scores RCC potentiel les plus fréquents

Scores RCC potentiel	Effectif	Score RCC min	Score RCC max	Écart max
28.5	8	23	27	4
26	7	16	24	8

Table 9
Variability in RCC scores for the most frequent potential RCC scores

Discussion

55 À partir de l'exemple de deux épreuves de raisonnement logique, nous avons illustré qu'il était possible d'identifier des différences de qualité dans les réponses fausses, d'ajouter alors des graduations à l'échelle de mesure et enfin d'identifier des différences interindividuelles, à même niveau global, en matière de qualité des erreurs commises dans ces tests. Le premier point, largement ignoré dans ce type de test, pourrait s'appliquer à différentes épreuves afin d'apporter des éléments d'information sur la nature des difficultés rencontrées dans l'épreuve. Cette approche diagnostique, déjà développée dans le cadre d'évaluation de connaissances (e.g. les travaux cités en introduction ainsi que les modélisations développées par King, Gardner, Zucker & Jorgensen, 2004), permettrait d'enrichir les éléments recueillis lors de la passation de tests cognitifs. Le deuxième point, directement lié au premier, aboutit à une mesure plus précise des capacités des personnes par l'ajout d'une graduation entre la mauvaise réponse et la bonne réponse. Dans les contextes cliniques ou éducatifs, cette possibilité de réponse intermédiaire apporte des informations utiles pour déterminer le niveau de maîtrise de la compétence évaluée. Le troisième point concerne les différences individuelles. Quelle interprétation peut-on donner à ces différences ? Elles pourraient être la trace de compétences latentes de raisonnement non prises en compte par les scores classiques des épreuves (car centrés uniquement sur les items réussis), la trace de la présence d'un « potentiel » de progression (d'où le choix de notre qualitatif). L'une des questions qui se posent alors est la suivante : pour rendre compte des capacités/ compétences de raisonnement d'une personne, quel est le meilleur indicateur : le score classique ou le score potentiel ? Cette question se pose de manière encore plus vive dans le cas des élèves qui présentent un score potentiel très supérieur à leur score initial. Il est donc nécessaire de compléter l'analyse a priori des erreurs, et l'approche descriptive présentée ici, par la recherche des processus mentaux qui peuvent expliquer ces erreurs. Des recherches doivent donc se poursuivre pour préciser la validité de ces indicateurs (scores RP et scores potentiel) ainsi que la fidélité des scores de gain. En attente, il faut donc considérer ces analyses et ces indicateurs, comme complémentaires aux scores classiques.

56 On pourrait relier notre approche au courant de l'évaluation dynamique qui vise, par d'autres références théoriques et d'autres méthodologies, à rendre compte également d'un potentiel d'apprentissage et/ou de développement (Chartier & Loarer, 2008 ; Hessels & Hessels-Schlatter, 2010 ; Huteau & Lautrey, 1999).Signalons que des études sont en cours qui confrontent les mêmes sujets aux deux situations (passation RCC et évaluation dynamique) afin de vérifier l'hypothèse de l'existence d'une relation entre ces deux types d'indicateurs de potentialités.

57 Reste que l'analyse des erreurs aux matrices se heurte, et nous l'avons signalé, au fait que la répartition des erreurs n'est pas équilibrée, car l'élaboration des distracteurs n'a pas été conçue à l'origine à partir de ce cadre. Par exemple, seuls 14 items sur 60 proposent une erreur de type 3, et la majorité des distracteurs est de type 5 (voir annexe 2). Il est ainsi plus difficile d'appliquer cette approche d'analyse de l'erreur sur un test existant que lors de l'élaboration d'une épreuve dans laquelle on peut sélectionner les distracteurs en fonction de critères prédéfinis. Comme d'ailleurs l'indiquent les auteurs du Raven, pour procéder à une analyse plus systématique des erreurs aux matrices, il serait

préférable de développer une nouvelle épreuve (Raven et al., section 1). Pour le test RCC, la typologie des auteurs proposée mériterait d'être affinée pour être adaptée à chaque item. On pourrait, par exemple, repérer des erreurs typiques liées à un item piégeant, ou à une confusion des familles, ou encore à une erreur de calcul. Utiliser une procédure de verbalisation des résolutions permettrait d'apporter des éléments d'information pertinents afin d'élaborer des modèles psychologiques explicatifs de la production d'erreurs.

58 Enfin, pour pouvoir interpréter les erreurs d'une personne, il conviendrait de pouvoir disposer d'éléments de comparaison, de normes, comme la distribution des erreurs pour des sujets de même niveau (ce que nous avons réalisé pour quelques scores dans l'étude présente) afin d'apporter une validation aux comparaisons envisagées.

59 Un autre point serait également à prendre en compte : les différences éventuelles de style de réponse. Certaines personnes préfèrent ne pas donner de réponse lorsqu'elles ne sont pas sûres qu'elle soit correcte, tandis que d'autres peuvent montrer moins d'hésitation et donner une réponse même si elles pensent qu'elle n'est pas juste, ou même peuvent répondre au hasard (ainsi, les analyses des enquêtes PISA montrent que les élèves français ont tendance à s'abstenir davantage à certains items : Pons, 2011). Il faudrait alors revoir les consignes de passation pour tenter de réduire ce biais possible ou, au minimum, aborder le style de réponse de la personne lors de l'entretien de restitution des résultats afin de nuancer, si nécessaire, ce score potentiel.

Conclusion

60 Bien que l'intérêt de l'analyse des erreurs dans certains tests fasse l'objet de controverses (pour les matrices de Raven : Grégoire, 2004), et conscients des limites de nos études qui restent exploratoires, nos résultats permettent cependant de poser quelques bases de réflexion sur la place possible de cette analyse dans les pratiques d'évaluation et de formation. L'erreur peut être ainsi une occasion de « réfléchir, de rectifier ou d'approfondir ses connaissances, de mieux se connaître soi-même » (Baruk, 1985, p. 73). Cette utilisation de l'erreur dans un cadre pédagogique comme moyen pour faire réfléchir les élèves sur leur propre fonctionnement intellectuel, pour faire évoluer la pensée est souvent évoquée (Bayet & Seknadjé-Askénazi, 1998 ; Favre, 1995). Elle peut également concerner les aspects conatifs des apprentissages comme l'indique Daniau : « s'intéresser aux erreurs commises, en habituant les élèves à les analyser en même temps que le maître, crée des conditions psychologiques favorables à l'apprentissage. Quand l'élève comprend que l'enseignant s'en préoccupe pour lui permettre de les dépasser, sans y attacher une valeur morale négative, la situation d'apprentissage se « dédramatise » et place l'enfant dans des conditions favorables d'assimilation » (1989, p. 133). Cette approche pourrait aussi être appliquée aux démarches d'entretien de restitution des résultats de tests ou d'entretien mené dans une perspective métacognitive. Cet aspect figure d'ailleurs dans le manuel des matrices : « bien que les tests des progressive matrices aient été construits pour fournir, aussi rapidement et simplement que possible, un indice général de la capacité d'une personne à percevoir et à raisonner clairement, nombre de psychologues les utilisent comme un outil d'identification des erreurs de raisonnement, et beaucoup de programmes d'apprentissage, notamment ceux de Jacobs et Feuerstein (dans ses Programmes d'enrichissement instrumental) se fondent sur une appréhension de la nature des erreurs » (Raven et al., section 3, 1998, p. 56). Analyser les

erreurs, tenter de les comprendre et de les modéliser, c'est aussi se donner les moyens pour développer des situations de remédiation adaptées. L'analyse des erreurs peut ainsi être la source de pratiques d'évaluation assez novatrices qui complètent une évaluation traditionnelle par un retour réflexif, avec le sujet, sur les items échoués afin de cerner la capacité de progression de la personne (Barbot, 2010).

61 Plus globalement, l'analyse des erreurs permet de recueillir des informations spécifiques sur les difficultés de la personne dans l'épreuve et apporte alors des éléments supplémentaires utilisables dans la phase de restitution des résultats, phase de l'évaluation qui prend de plus en plus d'importance actuellement, tout particulièrement dans le cadre du conseil en orientation (Aubret & Blanchard, 2005). Dans cet objectif, des rapprochements peuvent être développés avec d'autres perspectives de renouvellement des tests, comme celui de l'évaluation des stratégies de résolution (Rozencwajg, 2005) mais aussi, nous l'avons déjà indiqué, le courant de l'évaluation dynamique, approches qui, comme la nôtre, accordent également une place à d'autres indicateurs que le simple niveau de performance.

62 Malgré ses limites, l'approche que nous proposons nous apparaît donc comme une piste possible pour développer des méthodologies d'évaluation qui apporteraient des éléments d'information sur certains aspects du fonctionnement cognitif des individus dans les situations de résolution des tâches du test et ainsi permettraient, par leur caractère diagnostique, d'améliorer les procédures et outils d'évaluation disponibles actuellement.

BIBLIOGRAPHIE

Astolfi, J-P. (1997). L'erreur, un outil pour enseigner. Paris : ESF.

Aubret, J. et Blanchard, S. (2005). Pratique du bilan personnalisé. Paris : DUNOD.

Babcock, R-L (2002). Analysis of age differences in types of errors on the Raven's Advanced Progressive Matrices. Intelligence, 30, 485-503.

Barbot, H. (2010). Évaluation psychomé-trique des élèves en grande difficulté scolaire pour un accompagnement et un conseil en -orientation. Communication au Colloque international L'accompagnement à l'orientation aux différents âges de la vie, Inetop/CNAM, Paris, 17-19 mars.

Baruk, S. (1985). L'âge du capitaine. De l'erreur en mathématiques. Paris : Seuil.

Bayet, B. et Seknadjé-Askénazi, J. (1998). Les enjeux de l'évaluation. La nouvelle revue de l'AIS, 4, 131-136.

Bellina, I. (non daté). Style de raisonnement et progressives matrices de Raven. Communication personnelle.

Bonnardel, R. (1957). Nouvelle recherche sur un test différentiel de compréhension verbale abstraite. (BV16). Le travail humain, 3-4, 339-349.

Chartier, P. (2008). Expérimentation d'une épreuve de facteur g utilisant comme support des cartes à jouer. In E. Loarer, J-L. Mogenet, F. Cuisinier, H. Gottesdiener, P. Mallet et P. Vrignaud

(éd.), Perspectives différentielles en psychologie (pp. 39-42). Rennes : Presses Universitaires de Rennes.

Chartier, P. & Loarer, E. (2008). Évaluer l'intelligence logique. Paris : DUNOD.

Chartier, P. (2010). Prise en compte de différentes formes de variabilité interindividuelle dans des tests de type facteur g. In A., De Ribaupierre, P., Ghisletta, T., Lecerf et J.-L., Roulin (éd.), Identité et spécificité de la psychologie différentielle (pp. 81-86). Rennes : Presses Universitaires de Rennes.

Chartier, P. (2012). Évaluer les capacités de raisonnement avec les tests RCC. Paris : Eurotests Éditions.

Daniau, J. (1989). L'évaluation dynamique à l'école élémentaire. Paris : Armand Colin.

Favre, D. (1995). Conception de l'erreur et rupture épistémologique. Revue française de pédagogie, 111, 85-94.

Favre, D. (2010). Cessons de démotiver les élèves. Paris : DUNOD.

Fayol, M. (1995). La notion d'erreur, éléments pour une approche cognitive. In G. Blanchet, J. Raffier et R. Voyazopoulos (éds.), Intelligence, scolarité et réussite (pp. 137-152). Paris : La pensée sauvage.

Grégoire, J. (2004). L'examen clinique de l'intelligence de l'adulte. Sprimont : Mardaga.

Hessels, M.G.P. & Hessels-Schlatter, C. (2010, éds.). Évaluation et intervention auprès d'élèves en difficulté. Berne : Peter Lang.

Huteau, M. & Lautrey, J. (1999). Évaluer l'intelligence. Paris : Presses Universitaires de France.

King, K., Gardner, D., Zucker, S. & Jorgensen, M. (2004). The distractor rationale taxonomy: enhancing multiple-choice items in reading and mathématics. Pearson Education.

Lafontaine, D. & Raîche, G. (2011). Principes méthodologiques et techniques des enquêtes internationales, Mesure et évaluation en éducation, 34(2), 25-55.

Leplat, J. (1999). Analyse cognitive de l'erreur. Revue européenne de psychologie appliquée, 49(1), 31-41.

Loye, N. (2010). 2010, odyssée des modèles de classification diagnostique (MCD). Mesure et évaluation en éducation, 33(3), 75-98.

Pire, G. (1957). Test MGM. Manuel d'instructions. Paris : Éditions scientifiques et psychotechniques.

Pons, X. (2011). Les méthodes des enquêtes internationales et leurs fonctions politiques. L'exemple de la France face à PISA (1995-2008). Mesure et évaluation en éducation, 34(2), 57-85.

Raven, J-C., Court, J-H. & Raven, J. (1998). Manuel des Raven : section 3. Progressive matrices standard (PM38). Paris : EAP.

Raven, J., Raven, J-C. & Court, J-H. (1998). Manuel des Raven : section 1. Introduction générale aux tests de Raven. Paris : EAP.

Raven, J., Raven, J-C. & Court, J-H. (1998). Manuel des Raven : section 4. Les advanced progressive matrices. Paris : EAP.

Ribaupierre (de), A.,Ghisletta, P., Lecerf, T. et Roulin, J-L. (2010, éds.), Identité et spécificité de la psychologie différentielle. Rennes : Presses universitaires.

Thiébaut, E. (2000). Manuel du test BLS4. Paris : EAP.

Thiébaut, E.(2010). La notion de test différentiel : des hypothèses relatives aux régularités et différences de variabilité intra-individuelle. Illustration concernant l'épreuve BV16. In A. Ribaupierre (de), P. Ghisletta, T. Lecerf et J-L. Roulin (éd.), Identité et spécificité de la psychologie différentielle (pp. 141-145). Rennes : Presses Universitaires de Rennes.

Vodegel-Matzen, L., Van der Molen, M. et Dudink, A. (1994). Error analysis of Raven test performance. Personality and individual differences, 16(3), 433-445.

Wechsler, D. (2005). WISC-IV. Manuel d'administration et de cotation. Paris : ECPA.

NOTES

1. Par exemple, un total de 20 points dans une épreuve est considéré comme équivalent à tout autre total de 20, même si ces points ne sont pas obtenus sur les mêmes items.
2. Cette épreuve, qui vise à évaluer le raisonnement logique (de type facteur g), consiste à choisir une matrice qui vient compléter une configuration proposée.
3. Advanced progressive matrices.
4. Progressive matrices standard.
5. Nous convenons que cette valeur est assez arbitraire ici (voir plus loin ce qu'il en est pour le RCC), mais nous nous sommes inspirés de l'approche des échelles de Wechsler qui envisage également, nous l'avons rappelé, une cotation à trois niveaux.
6. Ici, la valeur de ½ point est moins arbitraire car on peut considérer que le point entier se décompose en ½ pour la valeur et ½ pour la famille.
7. Expliquable, en grande partie, par le fait que le score Raven potentiel repose sur le score Raven.
8. Même score RCC.

RÉSUMÉS

Dans les tests de raisonnement logique l'évaluation ne prend en compte le plus souvent que le seul score total. Pourtant d'autres variables mériteraient une analyse, comme les erreurs qui apportent des informations sur la nature des difficultés rencontrées par les sujets dans l'épreuve. À partir de l'identification de différentes qualités de réponses fausses, nous proposons le calcul d'indicateurs permettant d'observer des différences individuelles à ce niveau. Nous illustrons notre approche sur deux épreuves, les matrices de Raven et le test RCC, et des résultats d'élèves de 6e. L'intérêt de cette analyse fait l'objet de la discussion.

In the tests of logical reasoning, most often the assessment only considers the global score. Nevertheless, some other variables would be worth an analysis, such as the errors, which bring information about the nature of the difficulties the persons met with through the test. From the identification of different types of false answers, we propound to calculate indicators allowing to observe some individual differences on this level. Our approach is illustrated by two tests, Raven matrices and RCC, with sixth-form pupils' results. The interest of this analysis is the subject of the discussion.

INDEX

Keywords : erroranalysis, diagnostic assessment, Raven matrices, RCC
Mots-clés : analyse des erreurs, évaluation diagnostique, matrices de Raven, RCC

AUTEURS

PHILIPPE CHARTIER

est Maître de conférences en psychologie, INETOP/CNAM. Philippe Chartier est également responsable du groupe GEP (Groupe sur l'Evaluation des Personnes). Thèmes de recherche : méthodologie de l'évaluation, différences individuelles dans les processus de raisonnement, outils et méthodes d'accompagnement en orientation. Contact : Institut National d'Etude du Travail et d'Orientation Professionnelle, Centre de Recherche sur le Travail et le Développement, CNAM, 41, rue Gay-Lussac, 75005 Paris. Courriel : philippe.chartier@cnam.fr

HANA BARBOT

est Conseillère d'orientation-psychologue, Directrice de CIO (retraitée).. Hana Barbot était Conseillère d'Orientation Psychologue, directrice du CIO de Savigny sur Orge jusqu'en 2012. Thème de recherche : évaluation des élèves en grande difficultés scolaires et remédiations psychopédagogiques ; responsable académique du groupe de travail « mission des COP auprès des élèves handicapés ». Courriel : hana.barbot@orange.fr.

RODRIGUE OZENNE

est Conseiller d'orientation-psychologue, chargé de formation et de recherches, INETOP/CNAM. Thèmes de recherche : le bilan psychologique en orientation, l'entretien de restitution, les nouvelles technologies et orientation, le genre et la division sexuée en orientation. Institut National d'Etude du Travail et d'Orientation Professionnelle, Centre de Recherche sur le Travail et le Développement, CNAM, 41, rue Gay-Lussac, 75005 Paris. Courriel : rodrigue.ozenne@cnam.fr